Rate Matrix Prior (Draft)

Graham Jones

September 7, 2008

email: art@gjones.name
web site: www.indriid.com

1 Rate Matrix Prior

My convention is to have row vectors (of state frequencies) on the left acted on by transition matrices on the right. This seems to be the convention for Markov chains, although the opposite convention is generally more common. Rate matrices have rows summing to zero; transition matrices have rows summing to one. It is usual to impose the condition that the non-diagonal elements of a rate matrix sum to one, but I will work with unnormalised rate matrices.

For nucleotides, an arbitrary 12-parameter rate matrix, which I will call a non-time reversible, or NTR rate matrix, can be written as follows. Note that the numbering has w_i diagonally opposite to w_{i+6}. The diagonal entries follow from the fact that rows sum to zero.

To:	A	G	C	T
From				
A	-	w_1	w_2	w_3
G	w_7	-	w_4	w_5
C	w_8	w_{10}	-	w_6
T	w_9	w_{11}	w_{12}	-

Let $v_i = \log(w_i)$ for $1 \leq i \leq 12$.

The GTR rate matrix can be written as follows ([1] p205).

To:	A	G	C	T
From				
A	-	$\pi_G a$	$\pi_C b$	$\pi_T c$
G	$\pi_A a$	-	$\pi_C d$	$\pi_T e$
C	$\pi_A b$	$\pi_G d$	-	$\pi_T f$
T	$\pi_A c$	$\pi_G e$	$\pi_C f$	-

where $a, ...f$ are arbitrary positive numbers. Note that there are 10 parameters but there is also a redundancy, since if all of $\pi_A, \pi_G, \pi_C, \pi_T$ are multiplied by x and all of $a, ...f$ are multiplied by x^{-1} the same rate matrix is produced. The dimensionality of the unnormalised GTR rate matrix is therefore 9. Usually $\pi_A, \pi_G, \pi_C, \pi_T$ are normalised to sum to one, so that they can be interpreted as state frequencies at equilibrium, but only the form of the above GTR rate matrix is of concern here, not the meaning of the parameters. By comparing the NTR and GTR rate matrices, three conditions can be found relating ratios of the w_i, which become sums and differences of the v_i.

$$v_1 - v_7 + v_8 - v_2 + v_4 - v_{10} = 0 \tag{1}$$

$$v_7 - v_1 + v_2 - v_8 + v_{11} - v_5 + v_6 - v_{12} = 0 \tag{2}$$

$$v_7 - v_1 + v_3 - v_9 + v_{10} - v_4 + v_{12} - v_6 = 0 \tag{3}$$

The HKY rate matrix can be written as follows ([1] p201).

To:	A	G	C	T
From				
A	-	$\pi_G(a+b)$	$\pi_C b$	$\pi_T b$
G	$\pi_A(a+b)$	-	$\pi_C b$	$\pi_T b$
C	$\pi_A b$	$\pi_G b$	-	$\pi_T(a+b)$
T	$\pi_A b$	$\pi_G b$	$\pi_C(a+b)$	-

where $a = \alpha_R/\pi_R = \alpha_Y/\pi_Y$ in Felsenstein's notation. From this, four further conditions can be found, namely

$$v_8 = v_9, \qquad v_{10} = v_{11}, \qquad \text{and} \qquad v_2 = v_4 \tag{4}$$

$$v_7 - v_8 + v_{12} - v_4 = 0 \tag{5}$$

The two-parameter Kimura rate matrix ([1] p196), which I'll denote as KIM, is

To:	A	G	C	T
From				
A	-	b	a	a
G	b	-	a	a
C	a	a	-	b
T	a	a	b	-

From this, three further conditions can be found, namely

$$v_9 = v_{11}, \qquad v_2 = v_3, \qquad \text{and} \qquad v_4 = v_{10} \tag{6}$$

Equations (1) - (6) are linear constraints of form $\sum \lambda_i v_i = 0$. It can be shown that these 3+4+3=10 linear constraints are all independent. The four nested models therefore become four real vector spaces $V_{KIM} < V_{HKY} < V_{GTR} < V_{NTR}$ of dimensions 2, 5, 9 and 12. It is now straightforward to calculate the squared Euclidean distances d_{GTR} from an arbitrary rate matrix R to the nearest GTR rate matrix, d_{HKY} from that point in V_{GTR} to the nearest HKY rate matrix, and d_{KIM} from that point in V_{HKY} to the nearest KIM rate matrix. The squared Euclidean distance from R to the nearest HKY rate matrix is then $d_{GTR} + d_{HKY}$, and to the nearest KIM rate matrix it is $d_{GTR} + d_{HKY} + d_{KIM}$. By weighting the components, a prior such as $\exp(-(w_{GTR}d_{GTR} + w_{HKY}d_{HKY} + w_{KIM}d_{KIM}))$ can express a statement that a rate matrix R is likely to be very close to a GTR rate matrix, and probably quite close to a HKY rate matrix, and 'no further opinion' about closeness to a KIM rate matrix (by using large w_{GTR}, moderate $w_{HKY}, w_{KIM} = 0$) and similar statements.

I have spelled out the details for the particular models GTR, HKY, KIM because they are popular and have a neat mathematical structure. The general idea could of course be applied to any set of nested models, but the vector subspaces $V_{KIM}, V_{HKY}, V_{GTR}$ become arbitrary manifolds, and then calculating a distance to the nearest point in one could become more complex and the meaning of the distance harder to understand.

A prior like this removes the need to make a categorical decision about which substitution model is appropriate for a particular analysis, using simpler models for smaller data sets and so on. A Bayesian approach with a well designed prior should be able to deal with differing amounts of data automatically.

The strand-symmetric rate matrix is

To:	A	G	C	T
From				
A	-	a	b	c
G	f	-	d	e
C	e	d	-	f
T	c	b	a	-

Further details to do...

2 Independence of the constraints

This is a sort of appendix, and isn't very interesting.

2.1 Independence of the GTR constraints

From (1), (2), (3), all the v_i can be expressed in terms of these nine: $v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_9, v_{10}$ as follows. From (1)

$$v_8 = -v_1 + v_2 - v_4 + v_7 + v_{10} \tag{7}$$

From (3)

$$v_{12} = v_1 - v_3 + v_4 + v_6 - v_7 + v_9 - v_{10} \tag{8}$$

From (2)

$$v_{11} = v_1 - v_2 + v_5 - v_6 - v_7 + v_8 + v_{12}$$

Substituting for v_8 and v_{12} gives

$$v_{11} = v_1 - v_2 + v_5 - v_6 - v_7 - v_1 + v_2 - v_4 + v_7 + v_{10} + v_1 - v_3 + v_4 + v_6 - v_7 + v_9 - v_{10}$$

which simplifies to

$$v_{11} = v_1 - v_3 + v_5 - v_7 + v_9 \tag{9}$$

2.2 Independence of the HKY constraints

Assume (1), (2), (3), (4) and (5). Now all the v_i can be expressed in terms of these five: v_1, v_2, v_3, v_6, v_9 as follows. Adding (2) and (3) gives

$$v_5 - v_3 = 2v_7 - 2v_1 + v_2 - v_8 - v_9 + v_{10} - v_4 + v_{11} \tag{10}$$

Using (4) this can be written as

$$v_5 - v_3 = 2(v_7 - v_1 + v_2 - v_8 - v_4 + v_{10}) \tag{11}$$

and the right hand side is zero from (1). So

$$v_5 = v_3 \tag{12}$$

Also using $v_4 = v_2$ from (4) and (1),

$$v_7 - v_1 - v_8 + v_{10} = 0 \tag{13}$$

Subtracting (3) from (2) and using (4), (5) and (12) gives

$$v_2 - v_3 + v_6 - v_{12} = 0 \tag{14}$$

We have $v_4 = v_2$, $v_5 = v_3$, $v_8 = v_9$, and it remains to express v_7, v_{10}, v_{11}, and v_{12} in terms of v_1, v_2, v_3, v_6, v_9. From (14)

$$v_{12} = v_2 - v_3 + v_6 \tag{15}$$

Substituting this into (5) gives

$$v_7 = v_9 - v_3 + v_6 \tag{16}$$

From (13) and (15)

$$v_{10} = v_1 + v_3 - v_6 \tag{17}$$

and since

$$v_{10} = v_{11} \tag{18}$$

from (4), we are done.

2.3 Independence of the KIM constraints

Assume (1), (2), (3), (4), (5) and (6). Now all the v_i can be expressed in terms of v_1 and v_2 as follows.

Using (12), (18) along with (4), (5) and (6) it follows that

$$v_3 = v_4 = v_5 = v_8 = v_9 = v_{10} = v_{11} = v_2 \tag{19}$$

Now from (17) and (19) we have $v_7 = v_6$ and from (15) and (19) we have $v_{12} = v_6$. From (2) and (19) it then follows that $v_7 = v_1$, so

$$v_6 = v_7 = v_{12} = v_1 \tag{20}$$

References

[1] J. Felsenstein, *Inferring Phylogenies*, Sinauer Associates, Inc., 2004